Data Publication in Environmental Science

Project funded by DFG
http://umwelt.wikidora.com

Andreas Hense
andreas.hense@brsu.de

Florian Quadt
florian.quadt@brsu.de

CNI Spring 2012 Member Meeting
April 2-3, Baltimore
Topics of this Talk

- Why is data publication important?
- Which data do we deal with?
- Where are our data stored?
- How do we publish data?
The Data Publication Cycle

- **Experiment**
- **Interpret and annotate data**
- **Perform quality assurance**
- **Publish and expose**
- **Search and browse**
- **Inspect and learn**
- **Internalisation**
- **Externalisation**

[Hense and Quadt, 2011]
The Knowledge Spiral

[Hense and Quadt, 2011]
Differences between Text and Data Publication

<table>
<thead>
<tr>
<th></th>
<th>Text</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>File formats</td>
<td>few widely accepted file formats</td>
<td>many, disciplines, tools</td>
</tr>
<tr>
<td>Contents</td>
<td>optimised for human reader</td>
<td>sometimes difficult to inspect</td>
</tr>
<tr>
<td>SQA</td>
<td>peer-review long tradition</td>
<td>peer-review data??</td>
</tr>
<tr>
<td>FQA and TQA</td>
<td>typesetting, etc.: easy</td>
<td>partial automation possible</td>
</tr>
<tr>
<td>Browse&amp;Search</td>
<td>metadata plus full text search</td>
<td>often restricted to metadata</td>
</tr>
<tr>
<td>Storage site</td>
<td>single file, stored in a repository</td>
<td>links to huge files, data curation.</td>
</tr>
</tbody>
</table>

[Treloar et al., 2007]
[Klump, 2008]
How to Acquire High Quality Data?

**Reputation** If a repository has the reputation of accepting only high quality research data and the visibility of its publications is high, a publication in such a repository is of great value for the researcher.

**Reliability** If the organisation which is running the repository is known to be well-financed and has a long tradition, researchers will believe that their data are stored safely and for the long term. (Scientists love DOIs)

**Process** If the submission and quality assurance process is well-documented and user-friendly, it will engender trust and lower the workload for potential submitters. [Hense and Quadt, 2011]
Environmental Data
Observational Data

**Characteristics**  Small amount, heterogeneous.

**Exposition/Registry**  WDCC, GetInfo TIB (DOI).

**Provenance**  Convective and Orographically-induced Precipitation Study [COPS, 2007], General Observation Period [GOP, 2007].
Simulation Data

Characteristics  Huge amount, homogeneous.
Exposition/Registry  WDCC, GetInfo TIB (DOI).
Provenance  Coupled Model Intercomparison Project Phase 5 (CMIP5) [LLNL, 2010]
Data Repositories
... and examples

<table>
<thead>
<tr>
<th>institutional</th>
<th>domain-specific</th>
<th>general</th>
</tr>
</thead>
<tbody>
<tr>
<td>national</td>
<td>unusual</td>
<td>[ARROW, 2008]</td>
</tr>
<tr>
<td></td>
<td>[AuScope, 2009]</td>
<td>[TIB, 2010]</td>
</tr>
<tr>
<td></td>
<td>[IMOS, 2009]</td>
<td>[ANDS, 2009]</td>
</tr>
<tr>
<td>global</td>
<td>[DKRZ, 2010]</td>
<td>[DataCite, 2010]</td>
</tr>
</tbody>
</table>
The World Data Center for Climate (WDCC)

**Operator**  Deutsches Klimarechenzentrum GmbH: The German High Performance Computing Centre for Climate- and Earth System Research [DKRZ, 2010]

**Long-Term Archival Storage**  60 PB: 6 Sun StorageTek SL8500 tape libraries, 10 000 media slots per library, 8 robots per library.

**Annual Data Production**  Total: 50 PB, Limit for mass storage archive (expiration date): 10 PB, Limit for long-term data archive: 1 PB
Publication of Environmental Data
About the Project

**Funding**  
DFG, April 2009 – March 2012

**Partners**  
- Bonn-Rhine-Sieg University oAS
- Bonn University, Meteorological Institute
- German Climate Computing Centre (DKRZ)

**Objectives**  
- Definition of a *standard procedure* for the publication of data including documentation of quality assurance actions.
- Development of the *workflow system Atarrabi* (Basque: good weather spirit) for the researcher and the publication agent.
- *Integration* into the Word Data Center for Climate (WDCC).
- *Generalisation* of the defined processes for other environmental sciences.
Data Publication in Environmental Science

The Project
The Workflow
Atarrabi

Metadata database at WDCC Hamburg

1. Metadata import (stored procedures)
2. Scientist wizard
3. Agent wizard
4. Updated metadata export (Atarrabi XML file)
5. DataCite XML import (web service)
6. DOI registration (web service)

Scientist

Publication agent at WDCC Hamburg

DataCite
Registration service

A. Hense & F. Quadt
Data Publication in Environmental Science
Atarrabi imports all relevant metadata of the meteorological experiment to publish from the CERA2 database at WDCC Hamburg.

In the course of the ‘scientist wizard’ the researcher checks, changes and augments the metadata and performs and documents the scientific quality assurance (SQA).
3 A publication agent at WDCC Hamburg double-checks the results of the scientist wizard and performs the technical quality assurance (TQA).

4 If all checks are passed successfully, the updated metadata are exported back to the CERA2 database.
After the metadata in the CERA2 database have been updated, an XML file will be generated which is compliant to the DataCite XML schema for DOI registration. This file is imported by Atarrabi for the next step.

The XML file which has been imported in the step before is used to register DOI, metadata and landing page at DataCite.
Selected Concepts of the Atarrabi Publication System 1

**Web-based**  No client-side installation necessary. Atarrabi is hosted as a web-based application.

**Workspaces**  After login the user is directed to a role-dependent workspace, which focuses on the relevant information and functions.

**Wizards**  The metadata review process is structured like a software installation wizard. The individual wizard pages deal with certain aspects of the metadata (e.g. authors, spatial coverage, instruments, spatial coverage, quality). The scientist can leave the wizard anytime and continue at a later time.
Selected Concepts of the Atarrabi Publication System II

**List-of-Values**  Based on data already stored in the CERA2 metadata database at DKRZ Atarrabi provides search-as-you type lists or dropdown menus for persons, institutes, location, units, identifier and relation types.

**Workflow**  Using a workflow engine behind the scenes ensures a well-defined and easily changeable process flow.

**Adaptability**  Atarrabi is designed to be adaptable to other ...

- data domains,
- data sources, and
- process flows
How to get the source:

Atarrabi has been published under a GPL license at Sourceforge: http://sourceforge.net/projects/atarrabi/
Uptake

- Atarrabi releases so far:
  - Atarrabi 1.0 (January 2011): start of production with scientist wizard supporting CMIP5 data (climate simulation).
  - Atarrabi 1.5 (April 2011): support for observational data.
  - Atarrabi 2.0 (November 2011): wizard for agent review.
  - Atarrabi 2.1 (March 2012): automated identifier registration using the TIB web services.

- Up until now about 50 DOIs have been successfully registered with the support of Atarrabi.

- Very positive feedback from Atarrabi users.
Next Steps

- Maintenance of Atarrabi is taken over by DKRZ.
- Adaptation to other data repositories and other data domains requires new effort.
- Our vision: extend Atarrabi to support peer-review of data!
Please do not forget to take home a flyer.
Time for Questions

[Hense et al., 2009]
References 1

- **ANDS (2009).**
  About ANDS.

- **ARROW (2008).**
  Monash university ARROW repository | home.

- **AuScope (2009).**
  AuScope discovery portal.
References II

- **COPS (2007).**
  COPS - convective and orographically-induced precipitation study.

- **DataCite (2010).**
  Welcome to DataCite.
  http://datacite.org/.

- **DKRZ (2010).**
  Deutsches klimarechenzentrum GmbH.
  http://www.dkrz.de/.
GOP (2007).
GOP - general observation period.
http://gop.meteo.uni-koeln.de/gop/doku.php.

Acquiring high quality research data.
*D-Lib Magazine*, 17(1/2).

Publikation Umweltdaten.
http://umwelt.wikidora.com/wikidora/.
References IV


References V
